Selecting the Right Similarity-Scoring Matrix

William R. Pearson[1]

[1]University of Virginia School of Medicine, Charlottesville, Virginia

ABSTRACT

Protein sequence similarity searching programs like BLASTP, SSEARCH, and FASTA use scoring matrices that are designed to identify distant evolutionary relationships (BLOSUM62 for BLAST, BLOSUM50 for SSEARCH and FASTA). Different similarity scoring matrices are most effective at different evolutionary distances. "Deep" scoring matrices like BLOSUM62 and BLOSUM50 target alignments with 20% to 30% identity, while "shallow" scoring matrices (e.g., VTML10 to VTML80) target alignments that share 90% to 50% identity, reflecting much less evolutionary change. While "deep" matrices provide very sensitive similarity searches, they also require longer sequence alignments and can sometimes produce alignment overextension into nonhomologous regions. Shallower scoring matrices are more effective when searching for short protein domains, or when the goal is to limit the scope of the search to sequences that are likely to be orthologous between recently diverged organisms. Likewise, in DNA searches, the match and mismatch parameters set evolutionary look-back times and domain boundaries. In this unit, we will discuss the theoretical foundations that drive practical choices of protein and DNA similarity scoring matrices and gap penalties. Deep scoring matrices (BLOSUM62 and BLOSUM50) should be used for sensitive searches with full-length protein sequences, but short domains or restricted evolutionary look-back require shallower scoring matrices. *Curr. Protoc. Bioinform.* 43:3.5.1-3.5.9. © 2013 by John Wiley & Sons, Inc.

Keywords: similarity scoring matrices • PAM matrices • BLOSUM matrices • sequence alignment

SIMILARITY SEARCHING, HOMOLOGY, AND STATISTICAL SIGNIFICANCE

Protein similarity scoring matrices dramatically improve evolutionary look-back time because they capture amino acid substitution preferences that have emerged over evolutionary time. Amino acid changes can range from biochemically conservative, e.g., leucine to valine or arginine to lysine, to dramatically different, e.g., tryptophan to glycine. Amino acid scoring matrices capture this evolutionary information; conservative changes receive positive scores, while nonconservative changes will receive the largest negative scores. As a result, statistical expectation values (*E*() values) based on amino-acid similarity scores are far more sensitive than percent identity for finding homologs (*UNIT 3.1*).

In this unit, we provide a brief overview of the history of scoring matrices, the algebra used to calculate scoring matrices, and the important concepts of matrix information content and matrix target evolutionary distance. Because finding distantly related protein sequences is more challenging than finding closely related sequences, the BLOSUM62 matrix used by the BLAST programs and the BLOSUM50 matrix used by the FASTA programs are designed to identify distant homologs using long (typically full-length) sequences. Understanding the explicit or implicit evolutionary models used in similarity scoring matrices makes it much easier to choose the right scoring matrix. Generally, searches for short domains (or with shorter query sequences) require shallower scoring matrices. Likewise, shallow scoring matrices can be more effective at highlighting common orthologs when comparing proteins that have diverged in the past 100 to 500 million years. While deep scoring matrices are more effective in identifying distant relationships, deep scoring matrices can also contribute to homologous overextension when two closely related domains are embedded in

Current Protocols in Bioinformatics 3.5.1-3.5.9, October 2013
Published online October 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi0305s43
Copyright © 2013 John Wiley & Sons, Inc.

nonhomologous protein contexts. Using the appropriate scoring matrix can improve both search sensitivity and alignment accuracy.

AMINO ACID SUBSTITUTION MATRICES: HISTORY AND CLASSIFICATION

The earliest amino acid scoring matrices were based on amino acid properties or genetic code differences, but modern amino acid scoring matrices are based on empirical measurements of amino acid replacement frequencies from large sets of homologous sequences (Schwartz and Dayhoff, 1978). Empirical replacement frequency scoring matrices can be divided into two types: those with an explicit evolutionary model and the BLOSUM scoring matrices. Model-based scoring matrices include Dayhoff's original PAM series of matrices (Schwartz and Dayhoff, 1978), which were updated by Jones, Taylor, and Thornton (Jones et al., 1992). More recently, Gonnet (Gonnet et al., 1992) and Vingron and Mueller (VT and VTML; Mueller et al., 2002) developed model-based parameters using alignments between more distantly related proteins.

Model-based scoring matrices are appealing because they can be calculated for alignments at any evolutionary distance. Dayhoff's original PAM250 matrix was calculated based on 1572 observed mutations in 71 families of proteins with alignments that were more than 85% identical. The frequency of mutations was normalized for 1% change (99% identity), or PAM1, and then extrapolated to much longer evolutionary distances simply by multiplying the replacement frequency matrix. Thus, PAM10 corresponds to ∼90% identity, PAM30, ∼75% identity, PAM70, ∼55% identity, PAM120, ∼37% identity, and PAM250, ∼20% identity. Table 3.5.1 presents a more comprehensive set of scoring matrices and target percent identities. More recently, Vingron and Mueller described strategies for estimating replacement frequencies that use measurements from a broader range of evolutionary distances. However, evolutionary models assume that the model accurately describes

Table 3.5.1 Scoring Matrix Target Identity, Information Content, and Alignment Length[a]

Matrix	Gap penalty[b]	% Identity	Bits/ position	Random alignment length	50-bit length
SSEARCH version 36.3.6					
BLOSUM50[c]	10/2	25.3	0.21	160	238
BLOSUM62	11/1	28.9	0.40	86	125
VTML 160[c,d]	12/2	23.9	0.25	139	200
VTML 140	10/1	28.4	0.44	82	114
VTML 120	11/1	32.1	0.54	62	93
VTML 80	10/1	40.5	0.74	47	68
VTML 40	13/1	64.7	1.92	18	26
VTML 20	15/2	86.1	3.30	11	15
VTML 10	16/2	90.9	3.87	9	13
BLAST version 2.2.27+					
BLOSUM50[c]	13/2	29.4	0.39	85	128
BLOSUM62	11/1	29.6	0.41	82	122
BLOSUM80	10/1	32.0	0.48	69	104
PAM70	10/1	33.9	0.58	56	86
PAM30	9/1	45.9	0.90	34	56

[a]Median percent identity, bits per aligned position, alignment length, and alignment length required for a 50-bit score based on searches of 140 random sequences against 240,000 real protein sequences using the specified scoring matrix and gap penalties.

[b]Gap open/extend penalty, where the total penalty is $open + r \times extend$, where r is the number of residues in the gap. Thus, a 10/2 penalty produces a penalty of 12 for a one residue gap, 14 for two residues, etc.

[c]Scaled in 1/3-bit units; all other matrices are scaled in 1/2-bit units.

[d]As calculated according to Mueller et al. (2002).

replacement frequencies over long evolutionary times (Mueller et al., 2002).

In 1992, Steve and Jorja Henikoff described a direct approach to counting replacement frequencies at long evolutionary distances (Henikoff and Henikoff, 1992). The BLOSUM scoring matrices avoided the problem of extrapolating from PAM1 replacement frequencies by counting replacement frequencies directly with the BLOSUM series of matrices. Rather than relying on alignments of relatively closely related proteins, they identified conserved BLOCKS, or ungapped patches of conserved sequences, in sets of proteins that were potentially very distantly related. They then counted the amino acid replacements within these blocks, using a percent identity threshold to exclude closely and more moderately related sequences. In their description of the BLOSUM matrices, they showed that BLOSUM62 performed much more effectively than either the PAM120 (BLOSUM62 equivalent information content) or the PAM250 matrix (BLOSUM45 equivalent) for identifying distant homologs. BLOSUM62 was then incorporated as the default for the BLASTP (*UNIT 3.4*) program, while FASTA (*UNIT 3.9*) and SSEARCH (*UNIT 3.10*) switched to the BLOSUM50 matrix, which is more sensitive than BLOSUM62, but requires longer alignments.

THE ALGEBRA OF SIMILARITY SCORING (LOG-ODDS) MATRICES

Scoring Matrices as Odds Ratios

Similarity scoring matrices for local sequence alignments, which are rigorously calculated by the Smith-Waterman algorithm (Smith and Waterman, 1981) and heuristically calculated by BLASTP (Altschul et al., 1990; Altschul et al., 1997) or FASTA (Pearson and Lipman, 1988), require scoring matrices that produce negative values on average between random sequences. If the average or expected matrix score is positive, the alignment will extend to the ends of the sequences, and be global, rather than local. Dayhoff's initial PAM matrices were calculated as log-odds ratios, the logarithm of the ratio of the alignment frequency observed after a given evolutionary distance divided by the alignment frequency expected by chance:

$$\log\left(\frac{\text{frequency in homologs}}{\text{frequency by chance}}\right)$$

The Henikoffs used the same odds-ratio algebra when developing the BLOSUM matrices, but calculated their transition frequencies by counting the number of weighted changes in different blocks.

In 1991, Altschul published a seminal paper (Altschul, 1991) that showed that any scoring matrix appropriate for local alignments (one with a negative expected score) could be treated as a "log-odds" matrix of the form: $\lambda s_{i,j} = \log(q_{ij}/p_i p_j)$, where s_{ij} is the score given to the i,j alignment, q_{ij} is the replacement frequency for amino acid i to j, and the $p_i p_j$ term gives the expected frequency of two amino acids aligning by chance. The λ term is used to scale the matrix so that individual scores can be accurately represented with integers. Widely used scoring matrix values typically range from -10 to $+20$, reflecting λ scale factors of ln(2)/2—half-bit units used by BLOSUM62 and PAM120—or ln(2)/3—third-bit units used by BLOSUM50 and PAM250. For example, the BLOSUM62 score for aligning aspartic acid ("D") with itself is $+6$, and BLOSUM62 is scaled in 1/2-bit units, so a D:D alignment in related proteins is $6 = 2.0 \times \log_2(q_{D,D}/p_D p_D)$ or $2^3 = 8$ times more likely to occur because of homology than by chance. Likewise, the BLOSUM62 matrix assigns a D:L alignment a score of -4, which means that it is $2^2 = 4$ times more likely to occur by chance than in the homologous blocks aligned for BLOSUM62.

This ratio of homologous replacement frequency to chance alignment frequency explains why modern scoring matrices can give very different scores to identical residues. In the denominator, amino acids are not uniformly abundant (common amino acids like L, A, S, and G are found more than four times more frequently than rare amino acids like W, C, H, and M; see *APPENDIX 1A* for a table of the 1-letter amino acid codes), so common amino acids often have lower identity scores than rare ones. Likewise, amino acids are not uniformly mutable—A, S, and T change frequently over evolutionary time, while W and C change rarely. Thus, the highest identity score in the BLOSUM62 matrix (Fig. 3.5.1) is 11, corresponding to a W:W alignment, while A, I, L, S, and V get identity alignment scores of 4. Differences in identity scores, together with positive scores for nonidentity alignments between conserved amino acids, explain why sequence similarity scores are dramatically more sensitive than percent identity for inferring homology (see *UNIT 3.1*).

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  X
A  4
R -1  5
N -2  0  6
D -2 -2  1  6
C  0 -3 -3 -3  9
Q -1  1  0  0 -3  5
E -1  0  0  2 -4  2  5
G  0 -2  0 -1 -3 -2 -2  6
H -2  0  1 -1 -3  0  0 -2  8
I -1 -3 -3 -3 -1 -3 -3 -4 -3  4
L -1 -2 -3 -4 -1 -2 -3 -4 -3  2  4
K -1  2  0 -1 -3  1  1 -2 -1 -3 -2  5
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1  1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -2 -1  1  5
W -3 -3 -4 -4 -2 -2 -3 -2 -2 -3 -2 -3 -1  1 -4 -3 -2 11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3 -3 -1 -2 -2 -3 -3  3  1 -2  1 -1 -2 -2  0 -3 -1  4
X -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

Figure 3.5.1 The BLOSUM62 matrix. The BLOSUM62 matrix used by BLASTP, BLASTX, and TBLASTN is actually 23 × 23: 20 amino acids plus X (any amino acid), B (D or E), and Z (N or Q). Only the lower half of the symmetric matrix is shown to highlight the identity scores on the diagonal. The most positive value is 11 (W:W alignment); the most negative is −4 (found for many hydrophobic/hydrophilic and small/large replacements). The BLOSUM62 matrix is scaled in 1/2-bit units, so the W:W alignment of 11 is $2^{5.5} = 45$ times more common in homologous proteins than by chance. Weighted by amino acid abundance, the average similarity score is about −1 half-bits.

Matrix Information Content, Target Identity, and Alignment Length

In addition to generalizing scoring matrices as log-odds matrices, Altschul (1991) also showed that log-odds scoring matrices have an associated information content (relative entropy) or score per aligned position ("bits-per-position"). "Bits-per-position" can be used to estimate the number of aligned residues required to produce a statistically significant score. Shallow scoring matrices (e.g., PAM/VTML 10, PAM/VTML 20, or PAM/VTML 40) have higher information content than deep matrices (BLOSUM62, PAM250), which means that a shorter alignment (10 to 50 residues) can produce a more statistically significant score. At the same time, shallower matrices tend to produce higher identity alignments, because they give higher positive scores to identities and more negative scores to replacements (Table 3.5.1 and Fig. 3.5.2). For example, if an alignment needs a 50-bit score to be significant in a database search (*UNIT 3.1*), and the average bit score for BLOSUM62 is about 0.4 bits per aligned position (Table 3.5.1), then about 50/0.4 = 125 residues must be included in the alignment. In contrast, the VT20 matrix provides about 3.3 bits per aligned position, so even a 15-residue alignment can be significant. Thus, in a large-scale similarity search that needs a 50-bit score for statistical significance, domains shorter than 125 amino acids, or DNA exons shorter than 375 residues, often would not produce statistically significant scores with BLOSUM62, the default matrix used by BLAST, while exons shorter than 50 residues can easily be detected with VT20.

"Shallow" scoring matrices have more information content because they give more positive scores to identities and more negative scores to nonidentical replacements by varying the q_{ij} term in the log-odds matrices (the $p_i p_j$ values do not depend on evolutionary distance). From the evolutionary perspective, sequences that have diverged for less time, e.g., 10% to 20% change, will have more identical residues and fewer replacements simply because there has been less time for the sequences to change. Alternatively, sequences that have less than 25% identity because of a large amount of change will have many fewer identities and many more conservative replacements (PAM200 sequences will be less than 25% identical, on average). The numerical basis for this difference can be seen in Fig. 3.5.2, which compares parts of a "shallow" (VTML 20) and "deep" (BLOSUM62)

VTML 20

	A	R	N	D	C	Q	E
A	7						
R	-7	8					
N	-6	-5	8				
D	-6	-12	-1	8			
C	-3	-7	-8	-14	12		
Q	-5	-2	-4	-4	-13	9	
E	-5	-10	-5	-1	-14	-1	7

BLOSUM62

	A	R	N	D	C	Q	E
A	4						
R	-1	5					
N	-2	0	6				
D	-2	-2	1	6			
C	0	-3	-3	-3	9		
Q	-1	1	0	0	-3	5	
E	-1	0	0	2	-4	2	5

Figure 3.5.2 Comparison of a "shallow" (VTML 20) and "deep" (BLOSUM62) scoring matrix. Both matrices are scaled in 1/2-bits. For the small part of the matrices shown here, the VTML20 matrix produces an average 2.80 half-bit identity score, and an average −0.59 nonidentical score (weighted by amino-acid abundance). In contrast, BLOSUM62 produces 1.86 for identities but only −0.06 for nonidentities. Thus, VTML20 targets shorter, higher-identity alignments, because it penalizes nonidentities much more strongly.

matrix. Thus, in addition to differing in information content, scoring matrices have range of target percent identities and alignment lengths (Table 3.5.1). Shallower scoring matrices produce shorter, more identical alignments, because they give more negative scores to nonidentical aligned residues. "Deeper" scoring matrices produce longer alignments with lower percent identities because the penalty for a mismatch is much lower and more conservative nonidentities get positive scores.

In practice, the relationship between scoring matrix evolutionary distance, information content, percent identity, and alignment length suggests two reasons for changing from the BLOSUM62 and BLOSUM50 matrices used by BLASTP and SSEARCH/FASTA. First, one should change to a shallower matrix when looking for short alignments. We need a shallower scoring matrix for short domains, short exons, or short DNA reads because deep scoring matrices like BLOSUM62 do not have enough information content to produce significant scores. Short alignments require shallow scoring matrices.

One should also use a shallower scoring matrix when looking for orthologs—sequences that differ because of speciation events and are likely to share similar functions—between "relatively" closely related organisms (100 to 500 My). Protein sequence comparison algorithms are very sensitive; BLASTP and SSEARCH routinely find significant alignments between human and yeast (1.2 billion year divergence) and human and *E. coli* (>2.4 billion years). Because of this sensitivity, a mouse-human comparison often reports not only the orthologs (sequences that diverged at the primate/rodent split 80 million years ago), but also dozens of more distantly related paralogs that may have diverged 200 to 2000 million years ago. Mouse and human orthologs share about 83% amino acid identity; thus, for mammals, the VTML 20 matrix is expected to find all orthologs and paralogs that have diverged over the past 200 million years, but the matrix is much less likely to identify paralogs that share less than 40% sequence identity (divergence time > 1000 million years).

SCORING MATRICES AND GAP PENALTIES

While there is an intuitive mathematical explanation for pairwise similarity scores from the log-odds perspective, sensitive sequence alignments require both aligned residues and insertion or deletion gaps. Unfortunately, we do not have an analytical model for gap penalties and evolutionary distances. The default gap-penalties provided for BLASTP, SSEARCH, and FASTA were determined empirically (e.g., Pearson, 1991) with a focus on identifying distant homologs. In general, default gap penalties for BLASTP and SSEARCH/FASTA are set as low as possible; lower gap penalties would convert alignments from local to global, which would invalidate the statistical estimates. Thus, when considering whether to change gap penalties to improve search selectivity for a particular protein family, gap penalties should be increased (made more stringent), not decreased. Just as "shallower" scoring matrices target less divergence by giving higher scores to identities and more negative scores to nonidentities, gap penalties should increase with shallower scoring matrices (Reese and Pearson, 2002). Simulations to maximize the significance of short alignments suggest that for 1/2-bit scoring matrices, gap open penalties of $16.7-(0.067 \times \textit{pam-distance})$, e.g., $16.7-(0.067 \times 20) = 15$ for

VTML 20, and gap extend penalties of 2, are most effective (Reese and Pearson, 2002).

Low gap penalties can dramatically reduce the information content and average percent identity associated with a scoring matrix, and can dramatically increase the lengths of alignments produced by the matrix. The target percent identity, information content, and alignment lengths presented in Table 3.5.1 reflect the observed median values of the highest-scoring alignment produced by random queries against real protein sequences with the specified matrix and gap penalties. If gaps are not allowed, the average percent identity and information content increase and alignment length gets shorter. For example, if gaps are not allowed with BLOSUM62, the median percent identity increases from 28.9 (Table 3.5.1) to 33, information content almost doubles from 0.40 to 0.74, and median random alignment length drops from 86 to 45 residues. A similar effect is seen with VTML 80, where information content increases and alignment lengths decrease almost 2-fold when gaps are not allowed. Gap effects are less dramatic with shallower matrices like VTML 20—from 86% to 89% identity, from 3.3 to 3.5 bits per position, and from 11 to 10 residue median alignment lengths—because short evolutionary distances should allow many fewer insertions and deletions.

BLASTP Gap Penalties with Shallow Scoring Matrices

While the BLAST programs offer a set of scoring matrices with different evolutionary horizons (BLOSUM50 and BLOSUM62 are "deep"; PAM30 is relatively "shallow"), the modest gap penalties provided with their shallow matrices dramatically modify their effective evolutionary distance (Table 3.5.1). The "shallowest" combination of scoring matrix (PAM30) and gap penalties (9/1) requires an average of 56 aligned amino acids, or more than 160 nucleotides, to produce a 50-bit alignment score. Because these gap penalties are too low (Reese and Pearson, 2002), the BLAST protein matrices are less effective for short alignments or short evolutionary distances than they would be with higher penalties.

LONG ALIGNMENTS AND OVEREXTENSION

In addition to differing in information content (score or "bits" per aligned position) and optimal evolutionary distances (percent identity), different scoring matrices have different preferred alignment lengths (Table 3.5.1). Shallow scoring matrices have large negative values for amino acid replacements (Fig. 3.5.2), so alignments to nonhomologous (random) sequences will be short. Deep scoring matrices have less negative average replacement scores (VTML20's average nonidentity score is −5.8 half-bits, while BLOSUM62's is −1.2 half-bits), so their alignments tend to be longer. Table 3.5.1 (random alignment length column) summarizes the median alignment length between random queries and real protein sequences. BLAST and SSEARCH/FASTA statistics are very accurate (*UNIT 3.1*), so sequences that share statistically significant scores will always share a homologous domain. However, BLAST and SSEARCH/FASTA calculate *local* sequence alignments—the alignments begin and end at a position that maximizes the alignment score—so the boundaries of the alignment depend on both the location of the homologous domain *and* the scoring matrix used to produce the alignment. When a deep scoring matrix like BLOSUM62 is used to align more closely related sequences, the alignment can extend (overextend) into nonhomologous neighboring sequence. Gonzalez and Pearson (2010) termed this artifact "homologous overextension," and showed that it is a major source of errors in PSI-BLAST searches.

Homologous overextension often occurs from short repeated domains. For example, Figure 3.5.3A shows a BLASTP alignment of VAV_HUMAN (P15498) with SKAP2_XENTR (Q5FVW6), a protein that contains an SH3 domain that is homologous over 58 amino acids. However, the alignment is 198 residues long; the additional 140 residues in the alignment include a 100-residue Pleckstrin domain in SKAP2_XENTR that is not homologous (VAV_HUMAN contains an SH3 domain in the region that aligns to the Pleckstrin domain in SKAP2_XENTR). The 58-residue homologous SH3 domain contributes 85% of the bit score, with the additional 140 residues contributing less than 15% of the score. Using the slightly more stringent (shallower) BLOSUM80 matrix does not change the alignment overextension.

The FASTA programs offer a new option for identifying homologous overextension—subdomain scoring (Fig. 3.5.3B). By using the domain annotations available for one of the sequences to subdivide the alignment, it becomes apparent that the 58-residue SH3 domain is responsible for almost all of the

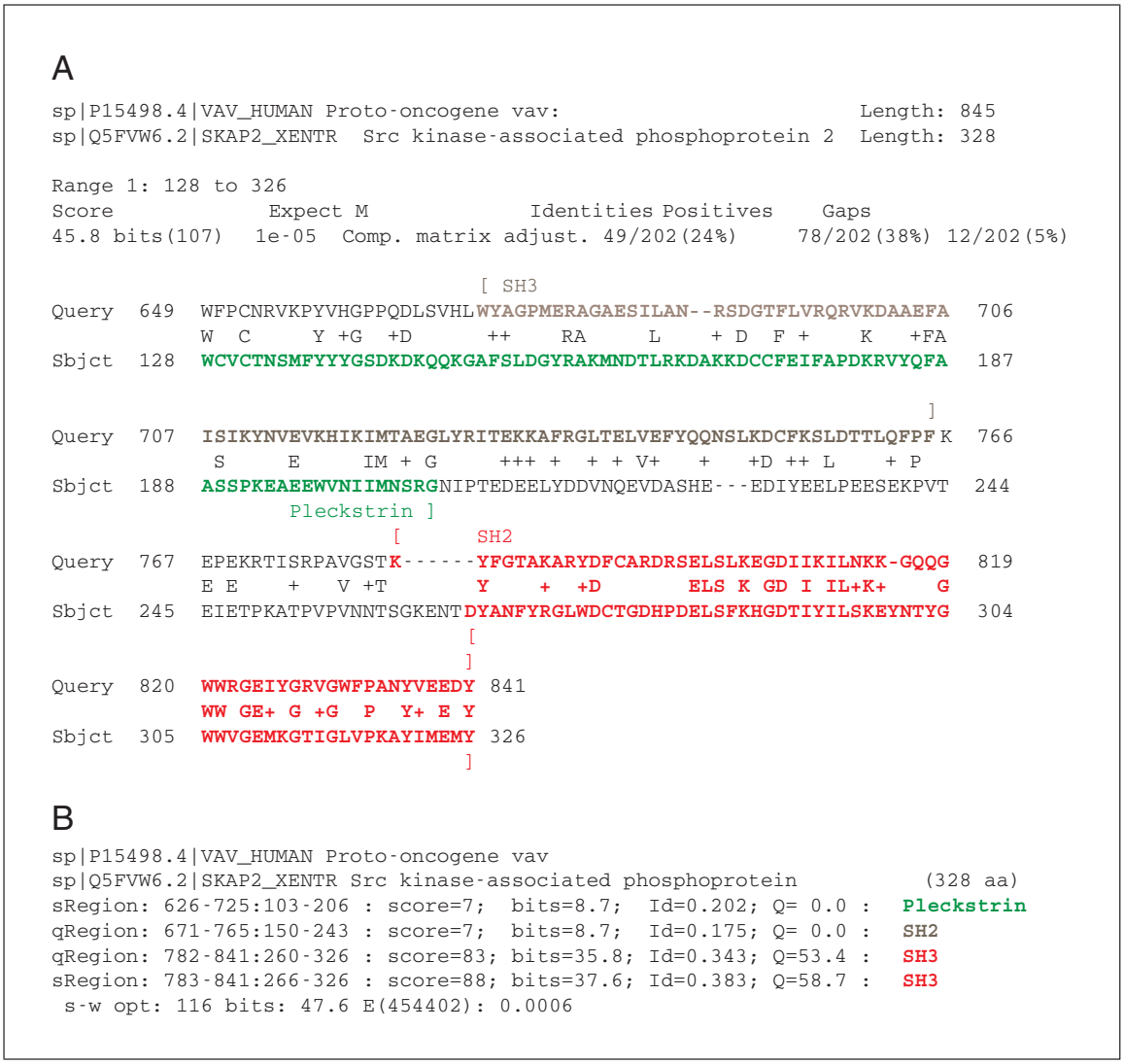

Figure 3.5.3 Overextension of an alignment of homologous SH2 domains. (**A**) BLASTP alignment of VAV_HUMAN with SKAP2_XENTR. The two proteins share a homologous SH2 domain (highlighted in red) over about 58 amino acids that contributes more than 85% of the similarity score. The remaining 140 amino acid alignment juxtaposes an SH3 domain from VAV_HUMAN (brown) with a Pleckstrin domain from SKAP2_XENTR (green). These two domains are not homologous; they are classified as having different folds in SCOP. (**B**) Sub-alignment scores produced by the SSEARCH36 program using the same scoring matrix as BLASTP (BLOSUM62, 11/1) for the VAV_HUMAN / SKAP2_XENTR alignment. Boundaries for annotated domains in the two proteins were taken from InterPro using the query VAV_HUMAN (qRegion) or the subject SKAP2_XENTR (sRegion). Thus, 103-206 for the Pleckstrin domain comes from InterPro annotations for SKAP2_XENTR, as does 671-765 for SH3 domain in VAV_HUMAN. The raw score, bit-score, and percent identity are shown for the subregions. The *Q*-score is $-10\log(p\text{-value})$ based on the bit score; thus $Q = 30$ corresponds to a probability (uncorrected for database size) of 0.001.

significant similarity found. It is often very difficult to judge the quality of a distant alignment visually; subdomain scoring provides a quantitative strategy for identifying overextension.

SCORING MATRICES FOR DNA

DNA scoring matrices, which are usually implemented as match/mismatch scores, can also be treated as log-odds matrices with target evolutionary distances (States et al., 1991). For example, the default match/mismatch penalties used by BLASTN in its most sensitive mode (`-task blastn`) uses a score of +2 for a match and −3 for a mismatch, which targets sequences at PAM10, or 90% identity (States et al. 1991). By default, searches on the NCBI nucleotide BLAST Web site use MEGABLAST (`-task megablast`), with match/mismatch scores of +1/−3 that target sequences that are 99% identical. By default, the FASTA program uses +5/−4 (also

available with BLASTN, `-task blastn`), which corresponds approximately to PAM 40, or 70% identity. Because DNA sequence comparison is much less sensitive than protein sequence comparison, it is very difficult to detect statistically significant DNA:DNA sequence similarity at distances greater than PAM 40 (PAM 40 is a short distance for protein comparisons).

In practice, the effective target identity for heuristic methods like BLAT, BLASTN, MEGABLAST, and other genome-alignment programs that do use scoring matrices, may be difficult to estimate from the reported match/mismatch scores. Heuristic programs typically use a hierarchy of filters to accelerate the similarity search, and each of those filters will affect the percentage identity and evolutionary distance of the alignments that are displayed. As a result, it is possible that the displayed alignments may have a lower percent identity than other possible alignments that were excluded during the early stages of the filtering process.

Ideally, the match/mismatch penalties used in genome alignment would match the evolutionary distances of the sequences being aligned; human DNA is expected to be more than 99.9% identical to itself, but human-mouse alignments in protein-coding regions will be less than 80% identical (outside of protein-coding regions, identity will typically be undetectable at <50%). Likewise, match/mismatch parameters should reflect potential alignment length; searches with short sequences will need higher match/mismatch ratios with higher information content (States et al., 1991).

SUMMARY

The BLAST and FASTA/SSEARCH protein-alignment programs use "deep" similarity scoring matrices like BLOSUM62 or BLOSUM50 to identify homologs that share less than 25% sequence identity. Deep scoring matrices require long sequence alignments to achieve statistically significant similarity scores and are more likely to extend alignments outside the homologous region. Shallower scoring matrices are more effective when searching for short homologous domains or short (<150-nt) exons, or when searching over shorter evolutionary distances. Scoring matrices that are matched to the evolutionary distance of the homologous sequences are also less likely to produce homologous overextension.

The match/mismatch ratios used in DNA similarity searches also have target evolutionary distances. The stringent match/mismatch ratios used by MEGABLAST are most effective at matching sequences that are essentially 100% identical, e.g., mRNA sequences to genomic exons. Deeper, more sensitive DNA scoring parameters are more effective for longer DNA evolutionary distances, e.g., mouse-human.

While scoring matrices and gap penalties can dramatically affect search sensitivity and alignment regions, modern sequence-comparison programs provide accurate similarity statistics, so it is unlikely that the wrong scoring matrix will produce a significant match to a nonhomologous protein. However, the wrong matrix can prevent short homologous regions from being found, or allow an overextension into a nonhomologous region from a homologous domain. The rapidly increasing volume of protein sequence means that close homologs will often be available, and shallower scoring matrices can produce more reliable, functionally informative alignments when closer homologs (>50% identical) are found.

ACKNOWLEDGMENTS

W.R.P. has been supported by funding from the National Library of Medicine (LM04969).

LITERATURE CITED

Altschul, S.F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555-565.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. A basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. 1992. Exhaustive matching of the entire protein sequence database. *Science* 256:1443-1445.

Gonzalez, M.W. and Pearson, W.R. 2010. Homologous over-extension: A challenge for iterative similarity searches. *Nucleic Acids Res.* 38:2177-2189.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* 89:10915-10919.

Jones, D.T., Taylor, W.R., and Thornton, J.M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* 8:275-282.

Mueller, T., Spang, R., and Vingron, M. 2002. Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.* 19:8-13.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11:635-650.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85:2444-2448.

Reese, J.T. and Pearson, W.R. 2002. Empirical determination of effective gap penalties for sequence comparison. *Bioinformatics* 18:1500-1507.

Schwartz, R.M. and Dayhoff, M. 1978. Matrices for detecting distant relationships. *In* Atlas of Protein Sequence and Structure, Volume 5, Supplement 3 (M. Dayhoff, ed.), pp. 353-358. National Biomedical Research Foundation, Silver Spring, Maryland.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

States, D.J., Gish, W., and Altschul, S.F. 1991. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *Methods Enzymol.* 3:66-70.